Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®

Artificial Intelligence 151 (2003) 241–245

Artificial Intelligence

www.elsevier.com/locate/artint

Response

Reply to Carruthers and Akman

Drew McDermott

Department of Computer Science, Yale University, PO Box 208285, New Haven, CT 06520-8285, USA

I am grateful to Professors Carruthers and Akman for their kind words about my book. I use this reply to clarify a couple of points, starting with Carruthers's review.

After surveying *Mind and Mechanism*, Carruthers concentrates on the details of my theory of phenomenal consciousness. His main criticism is this: "McDermott advertises himself as defending a higher-order theory of consciousness, maintaining (very roughly) that a phenomenally conscious sensation will be one of which the subject is aware. But in fact, many of the considerations he introduces in developing and defending his theory are purely first-order in character". It may be that my use of the phrase "higher-order" does not correspond exactly to the meaning intended by most consciousness theorists. But one thing the phrase really can't mean is that conscious sensations are those "of which the subject is aware". There are two problems with the quoted clause: the first is that "the subject" is part of what we have to explain; the second is that "is aware" comes too close to begging the question. We just can't explain my sensations in terms of what "I" am aware of, because the ultimate explanation will have to account for what this "I" is. It seems much more likely to me that the subject is explained in terms of consciousness than the other way around.

My understanding of the phrase "higher order" is that it explains consciousness in terms of beliefs about events in the nervous system (or other hardware) of a computational entity. The beliefs are higher-order because they include beliefs about beliefs (as well as beliefs about other things). The key question then is, *Whose* beliefs are we talking about? The proposed answer is that they are the beliefs of a module of the computational entity. To be as clear as I can: These beliefs are themselves purely computational entities that one can visualize as symbol structures being written, copied, and erased the way symbol structures are written, copied, and erased in computers. If that seems too abstract, picture symbols being written on pieces of paper by clerks. In this sense they are purely physical entities, but they also *mean something* because of the causal relationships they bear to the objects they are about.

E-mail address: drew.mcdermott@yale.edu (D. McDermott).

0004-3702/$ – see front matter
doi:10.1016/S0004-3702(03)00121-8

Carruthers argues that

> [McDermott] . . . never clearly says whether he believes that it is higher-order *thoughts* about our perceptual states which renders them phenomenally conscious, or whether we need instead to have higher-order *perceptions* of those states. He does often use the language of perception. For example, he talks about the "ability to perceive the output of sensory systems" (p. 106; cf. p. 133). On other occasions he seems to suggest that *any* sort of higher-order representation of perceptual states would do (p. 126). And on yet others he uses language strongly suggesting that it is a higher-order conceptualized *belief* which he has in mind. . . . This lack of clarity is unfortunate.

I hope it is clear from my gloss that I couldn't possibly be arguing for "thoughts" about perceptual states to be a crucial factor; I have no idea what a thought is, unless it can be reduced to some sort of computational entity. As I said above, "belief" is the best term I can think of, but words like "belief" and "perception" are extremely easy to misinterpret. The problem is that almost all our language about computational entities is metaphorical. We use words like "instruction" and "memory" to talk about how computers work, but they're not literally accurate. Unfortunately, they're very hard to do without. So when one attributes belief to a hypothesized module in the brain or in some future computer, it must be kept in mind that this is *not* full-blown belief in the sense we would use in talking about a person, but is instead some pale image that can be fully explained in terms of symbols being written, copied, and erased.

I have now explained what I mean by "higher-order", but unfortunately I am not sure what is meant by the term "first-order". In the book, I refer to "first-order theories" briefly as a straw man, an obviously inadequate theoretical construct. First-order theories "are those in which in some contexts the processing of sensory information is 'experience-like' in a way that allows us to say that in those contexts the processing *is* experience" (*M&M*, p. 22).

Carruthers says that modeling oneself as a physical entity does not require any higher-order representation.

> ". . . in predicting the likely movements of perceived objects [animals and robots] have to factor in the effects of their own intentional movements. But none of this requires any higher-order representation of the agent's own mental states. For example, in order to predict how a moving ball will change trajectory given that I have formed an intention to kick it, I don't have to represent my own intention qua mental state. Rather, my physics module just has to receive as input a representation of the intended movement of my leg in relation to the moving ball. And that can surely be generated from my intention without anything explicitly representing it as an intention."

I think the sense of "first-order" access to an intention here is the same as my term "normal" access, as opposed to "introspective" access (p. 108 of *M&M*). It's no doubt true that in the heat of a soccer game a good player rarely has time or motive for introspective access to her intentions; she kicks, then runs in the appropriate direction. There is also reason to doubt that a good player is consciously aware of all the events occurring in a fast-moving

situation. I realize this may sound strange. But just keep in mind that it is not hard to build robots that compute predicted ball trajectories from kicks (or good kicks from desired ball trajectories). No one supposes they are conscious.

When *I* kick a moving ball, there is a vast gulf between what I intend and the direction the ball goes. Physics tells me something about why the ball went over my head and not forward, but only in the context of my observation that there was an intention for the ball to go forward. As has frequently been noted, conscious intentions are sometimes less effective than unconscious ones, because the latter are often the result of training.

The point I want to make is that in higher-order theories there is no place for "first-order" properties of experience. It seems that Carruthers wants there to be a layer of actual experience that underlies judgments about experience, and perhaps this underlying layer is where "first-order" considerations arise. But such an underlying layer is exactly what I reject. The stuff that underlies judgments about experience is not a different, lower-order kind of experience; it's not experience at all, but computation. Experience is nothing but a particular class of judgment (themselves computational) about computations.

I agree with Carruthers that in the following quote (from p. 215), I overstated the case: "it seems as if any computational entity that dealt with a physical environment that included its own body would have to have a model of itself as a perceiver and decision maker; and in that model the entity and events involving it would have to be labeled as having the features of phenomenal consciousness". I should have said "any *intelligent* computational entity". But in context the overstatement doesn't seem too unforgivable. If we add in the last clause of the sentence before it, we get: "... we can expect consciousness to be a necessary component of a computational intelligence, not an inexplicable accident. In particular, it seems as if any computational entity ...".

Carruthers may be right that higher-order representations are not as common as I hypothesize. Perhaps it is mere sentimentality that leads me to conclude that apes and three-year-old humans can think about their own intentions and perceptions, and therefore are conscious. The fact that neither group can *talk* about their perceptions is probably irrelevant, although it certainly complicates investigation into the issue. However, there is no reason to think it is impossible ever to decide whether apes and toddlers have beliefs about their beliefs; it's an empirical question which will eventually get answered.

Another question concerns the fallibility of introspective judgments. If a sensation is a real-time observation of a perceptual event, then the observation could be erroneous. It seems as if I could think I was in pain and not actually be in pain, but that outcome is pretty unsatisfactory, so I tentatively concluded that such gaps would just erase themselves. The conclusion was tentative, and I am open to a different resolution of the problem. Carruthers evaluation is that, "If ... McDermott's view [is] that phenomenal consciousness exists whenever there is a higher-order belief to the effect that a perceptual experience is occurring, whether or not any such experience is really present[,] then it is surely unacceptable. When I undergo a phenomenally conscious experience of red I don't just find myself judging, blindly, that such an experience is taking place. Rather, if I make such a judgment it will be grounded in a fine-grained awareness of the character of my perceptual state". This is another instance of the confusion I mentioned at the outset, between "I" and the entity that is doing the higher-order believing. *Of course* when "I" undergo a conscious experience I don't find myself "blindly" judging anything. My brain's theory of the skill

of "I" in discriminating its experience includes firm beliefs that such experiences are fine-grained and immediate. I do occasionally find myself making "blind" judgments, as when I assume that if certain political leaders are in favor of something it must be a bad idea. It's in the nature of political judgments that they can seem blind even to their possessors, although we usually try to find some actual rationale for them when we realize how blind they are. It's in the nature of judgments about sensations that they never seem blind to their possessor. But all the beliefs in the self-model *are* blind, in the sense that they are simply arrived at by computational processes. It just happens that beliefs about sensations come equipped with an intuition that there is no blindness going on here. I don't find myself believing in some abstract sense that I am in pain; I find myself believing in a particular sort of pain with (ouch) *this* phenomenal quality. It does seem weird that such beliefs might have causes other than the usual causes of pains, so that the strong feeling that one's toe has been stubbed might be due to noise in a neuron between the toe and the brain. We have to say *something* about such cases. I would say that such beliefs are self-fulfilling; any belief by the self-model that has the same content as a belief normally caused by a stubbed toe *is* a sensation of a stubbed toe. I agree that it is not obvious that this is the correct thing to say.

About Professor Akman's review I have less to say, because he disagrees with me less. I do have a couple of clarifying remarks.

He glosses my theory of free will as depending on the infinite regress that would result if a decision-making robot tried to use a causal model of itself in deciding what to do. There is a sort of infinite regress here, but it's easy to misclassify it, as Akman seems to do in his footnote relating it to Tarski's hierarchy of metalanguages. It might seem as if my point is that no computer program could "understand itself" or "model itself" completely, lest it fall into Gödelish paradoxes. Actually, I take no position on such questions. It certainly seems as if a robot could, without any danger of paradox, inspect its own code and hardware in order to answer many questions about its behavior. For example, a robot could answer questions such as "How fast could I react to a speeding SUV that appeared while I was crossing the street?"; or "What percentage of the time will I go right rather than left when trying to get around an obstacle?" But one sort of question for which inspecting one's own program is useless is "What should I do *right now*?" It's uselessness does not depend on any tricky Gödel-style argument, but simply on the observation that the text of the program can reveal only that at some point the robot will consult the text of the program. The program is a perfect (if partial) causal model of how the robot works and how its decisions influence objects in its vicinity, including its own body. But this particular causal model should never be invoked when making a decision, because it is useless for that purpose. Hence in the overall world model, the robot must mark itself as exempt from causal modeling; as "free", in other words.

I must disagree with, or at least amend, the following paragraph from Akman's review:

> It is true that this computationalist explanation of consciousness has all the characteristics of "explaining away" rather than a true explanation. But, this is expected. It is, after all, hard to see how a computational model, incorporating inputs, outputs, and—what else—computations, can have phenomenal consciousness. One's best bet in this case is to argue, as McDermott does, that something *like* consciousness will be exhibited by the system.

This is in fact a tactical objective of the argument in *Mind and Mechanism*, and not the end point. I spend most of Chapter 3 talking about the sorts of self-models intelligent robots would probably have, and conclude that such robots would exhibit *virtual consciousness*, defined as behavior and beliefs that are identical to consciousness from an outside observer's viewpoint. On p. 131 I observe that "Virtual consciousness is the dependence on a self-model in which perceptions and emotions have qualia, some states of affairs are intrinsically better than others, and decisions are exempt from causal laws. . . . The question of whether a machine or organism [exhibits virtual consciousness] is purely a matter of third-person observation. . . . Testing whether a system exhibits *virtual* consciousness will eventually be completely uncontroversial, or at least only as controversial as testing whether a system has a belief".

However, in the next paragraph I argue that real consciousness and virtual consciousness will turn out to be the same thing. (Just as heat and random molecular motion are the same thing.) So my explanation of consciousness has the "flavor" of explaining it away, but I really do mean to explain it. The conclusion that a garden-variety property such as virtual consciousness is the same as the strikingly singular property of phenomenal consciousness may always seem counterintuitive, but it could turn out to be true nonetheless.